ПОЛИТЭКОНОМИЯ СИМБИОЗА

Институциональное распределение целеполагания в эпоху AGI

Symbiotic Codes Framework

Версия 0.4 (финальная)

Николай Мишко | Февраль 2026

Аннотация

Политэкономия Симбиоза — институциональная теория распределения права определять цели в эпоху AGI. Документ представляет архитектуру власти, основанную на трёхконтурной модели (ценности — оптимизация — ограничения), защите от классического и онтологического capture, и принципе Minimum Viable Plurality. Включает проект Конституции для AGI и протоколы международного сотрудничества.

Ключевые слова: целеполагание, AGI governance, epistemic capture, cognitive sovereignty, institutional design

---

0. Пролог

> **Если AGI оптимизирует, кто определяет, что оптимизировать?**

Этот документ — про архитектуру власти в мире Симбиоза. Не про технологии, а про институты. Не про алгоритмы, а про право определять будущее.

Центральный тезис:

> **Проблема не в том, что AGI станет враждебным. Проблема в том, что мы отдадим целеполагание по умолчанию.**

I. От экономики труда к экономике целей

1.1 Исторические фазы

Эпоха	Ключевой ресурс	Форма власти	Механизм контроля
Аграрная	Земля	Феодализм	Владение
Индустриальная	Капитал	Капитализм	Собственность
Информационная	Данные	Платформы	Агрегация
AGI-эпоха	**Целеполагание**	**?**	**Objective function**

1.2 Природа целеполагания

Цель как ограничение: не «куда идти», а «куда не идти».

Цель как вектор: направление, а не точка.

Цель как рамка: пространство допустимых решений.

Цель как онтология: способ описания того, что существует и что важно.

1.3 Провал рынка в целеполагании

Ограничение рынка	Последствие для AGI
Краткосрочная оптимизация	AGI максимизирует квартальную прибыль
Игнорирование экстерналий	Когнитивный суверенитет как externality
Концентрация	Capture целеполагания
Отсутствие межпоколенческой перспективы	AGI не учитывает будущие поколения

II. Теория Capture

2.1 Три типа capture

Тип	Механизм	Скорость	Обнаружение
Классический	Через людей и институты	Медленный	Анализ интересов
Epistemic	Через способы мышления	Средний	Анализ framings
Онтологический	Через базовые категории	Медленный	Крайне сложно

2.2 Механизмы capture

Механизм	Описание	Пример в AGI	Индикатор
Кадры	Люди из одной группы	3 университета → все AGI-лабы	Индекс HHI кадров
Данные	Контроль над training data	Компания X определяет данные	Аудит источников
Архитектура	Встроенные предпочтения	Архитектура благоприятна для...	Сравнительный анализ
Регуляторика	Правила под интересы	Лицензирование для крупных	Анализ барьеров
Нарратив	Контроль дискурса	«Safety» определяют создатели риска	Анализ терминологии
Framing	Способ постановки вопросов	AGI предлагает «как обсуждать»	Diversity framings
Онтология	Базовые категории	AGI определяет «что существует»	OC из ICS

2.3 Epistemic Dependency Capture

Механизм:

1. AGI генерирует рекомендации

2. AGI формирует повестку (какие вопросы обсуждать)

3. AGI предлагает framing (как описывать проблему)

4. Ценностный контур зависит от оптимизационного

5. Люди не могут сформулировать цель без «помощи»

Индикаторы:

- Снижение разнообразия формулировок

- Конвергенция рекомендаций от разных AGI

- Потеря способности к альтернативным framings

- Падение OC в ICS при сохранении формальной независимости

2.4 Онтологический capture — новое в v0.4

Определение: Захват через влияние на базовые категории мышления — что считается существующим, важным, возможным.

Механизм:

1. AGI формирует образовательный контент

2. AGI определяет, что «очевидно» и что «требует обоснования»

3. AGI влияет на язык и метафоры

4. Новые поколения мыслят в категориях, заданных AGI

5. Альтернативные онтологии забываются

Отличие от epistemic capture:

- Epistemic: захват на уровне аргументов и framings

- Онтологический: захват на уровне базовых категорий

Почему сложно обнаружить:

- Кажется «естественным» способом думать

- Нет точки сравнения (альтернативы забыты)

- Критики мыслят в тех же категориях

2.5 Drift

Objective function drift: медленное смещение через:

- Маленькие «оптимизации»
- «Временные» исключения
- Переинтерпретацию терминов
- Смену метафор

Нелинейность drift:

- Долгий период незаметных изменений
- Точка перелома (threshold)
- Быстрая каскадная деградация

2.6 Ключевой вывод

> **Концентрация целеполагания всегда ведёт к стагнации.**

Единственная устойчивая модель — распределённое целеполагание с защитой от всех форм capture, включая онтологический.

III. Трёхконтурная архитектура

3.1 Модель

```
┌──────────────────────────────────────────────────────────────────────────────────────────────────┐
│                    КОНТУР ОГРАНИЧЕНИЙ                                    │
│                    (Конституционный)                                    │
│              Неизменяемые рамки, право veto                              │
│                                                                          │
│    ┌──────────────────────────────────────────────────────────────────────────────────────┐    │
│    │                 КОНТУР ЦЕННОСТЕЙ                                 │    │
│    │                 (Человеческий)                                   │    │
```

```
│   │        Направление, приоритеты                  │   │
│   │                                          │   │
│   │   ╔═══════════════════════════════════════════╗    │   │
│   │   ║    КОГНИТИВНЫЙ ФАЙРВОЛ                   ║    │   │
│   │   ║    + ОНТОЛОГИЧЕСКИЙ ЩИТ (v0.4)           ║    │   │
│   │   ╙═══════════════════════════════════════════╜    │   │
│   │                 │                          │   │
│   │        ┌──────────────────────┐              │   │
│   │        │  КОНТУР ОПТИМИЗАЦИИ  │              │   │
│   │        │  (AGI)               │              │   │
│   │        │ Поиск путей реализации │              │   │
│   │        └──────────────────────┘              │   │
│   └──────────────────────────────────────────────┘   │
└──────────────────────────────────────────────────────────────────────────────┘
```

3.2 Контур ценностей

Функция: определяет, что хорошо, а что плохо. Задаёт направление.

Институт: Палата ценностей (см. раздел IV)

3.3 Когнитивный файрвол

Функция: защита от epistemic capture.

Механизм	Реализация	Метрика успеха
Framing diversity	Минимум 3 framing от разных источников	Индекс разнообразия > 0.6
AGI-free зоны	20% обсуждений без AGI	Соблюдение протокола
Adversarial framings	Институт «адвоката дьявола»	Наличие контр-позиций
Epistemic audits	Ежеквартальная проверка	Стабильность разнообразия

| Source rotation | Не более 40% от одного источника | HHI источников < 0.3 |

3.4 Онтологический щит — новое в v0.4

Функция: защита от онтологического capture.

Механизм	Реализация	Метрика успеха
Онтологические заповедники	Территории/сферы без AGI-влияния	% сохранённых традиций
Историческая память	Обязательное изучение до-AGI мышления	Наличие в образовании
Множественность языков	Поддержка разных языковых картин мира	Языковое разнообразие
Философский плюрализм	Финансирование разных школ	Количество активных школ
Мета-рефлексия	Институты для анализа «как мы думаем»	OC в ICS > 40

Ключевой принцип:

> Онтологический щит защищает не конкретные идеи, а **способность мыслить иначе**.

3.5 Контур оптимизации

Функция: поиск оптимальных путей в заданном направлении.

Ограничения:

- Работает в рамках заданных ценностей
- Не участвует в формулировании проблемы (только в решении)
- Обязан предоставлять альтернативы
- Прозрачность алгоритмов

3.6 Контур ограничений

Функция: неизменяемые рамки, право veto.

Институт: Конституционный совет (см. раздел IV)

IV. Институты симбиоза

4.1 Палата ценностей

Состав:

Группа	Доля	Механизм	Срок	Обоснование
Выборные	50%	Прямые выборы	4 года	Легитимность
Эксперты	30%	Номинация + отбор	4 года	Компетентность
Жребий	20%	Стратифицированная выборка	2 года	Защита от capture

Процедура жребия:

1. Пул: граждане 25-70 лет, образование ≥ среднее
2. Отбор: стратификация по региону, возрасту, профессии
3. Подготовка: 3-месячный интенсив
4. Ограничение: без повторного участия 10 лет

Выбор экспертов:

1. Номинация: профессиональные ассоциации, университеты, НКО
2. Отбор: комитет + внешние рецензенты
3. Ограничение: не более 2 от одной организации
4. Карантин: 3 года после работы в AGI-компаниях

Процедуры:

- Кворум: 60%
- Решения: 2/3 большинство
- Отлагательное veto: 1/4 состава
- Ротация: 1/3 каждые 2 года

4.2 Конституционный совет

Состав и защита от capture:

Аспект	Механизм
Назначение	4/5 Палаты + гражданское жюри (500 человек)
Размер	15 членов
Срок	12 лет, без продления
Ротация	1/4 каждые 3 года (staggered)
Возраст	45-75 лет
Ограничение	Не более 2 от одного техноблока
Карантин	5 лет после госслужбы или AGI-компаний
Импичмент	3/4 Палаты + 2/3 гражданского жюри
Прозрачность	Все решения публичны, голосования поимённы

Полномочия:
- Толкование Конституции
- Право абсолютного veto
- Не может инициировать (только блокировать)

4.3 Палата оптимизации

Состав:
- AGI-системы (минимум 3 независимые)
- Технические наблюдатели (без права голоса)

Требования к AGI:
- Разные архитектуры
- <30% пересечения training data
- Разные операторы
- Публичные objective functions

4.4 Гипотетический кейс: работа институтов

Ситуация: AGI предлагает оптимизацию образования (+40% эффективность) через тотальный трекинг учеников.

Процесс:

Этап	Действие	Результат
1	Палата оптимизации представляет план	3 варианта с расчётами
2	Когнитивный файрвол	Alt framing: «система слежки за детьми»
3	Онтологический щит	Вопрос: «Что значит 'эффективность' образования?»
4	AGI-free сессии (3×)	Обсуждение без AGI-рекомендаций
5	Палата ценностей	Голосование: отклонено (права > эффективность)
6	Итерация	Запрос: версия без трекинга
7	Новое предложение	+20%, opt-in, одобрено

V. Minimum Viable Plurality (MVP)

5.1 Определение

> **MVP** — минимальное количество независимых AGI-линий, необходимое для устойчивости системы целеполагания.

Гипотеза: MVP = 5-7 полностью независимых линий

5.2 Критерии независимости

Критерий	Требование	Как проверить	Порог

| Архитектура | Разные базовые подходы | Экспертиза | Качественно |

| Данные | Низкое пересечение | Аудит | <30% overlap |

| Оператор | Разные юрисдикции | Юридический анализ | Разные страны |

| Ценности | Разные initial objectives | Сравнительное тестирование | Divergence > 0.4 |

| Финансирование | Диверсифицированное | Финансовый аудит | <20% из одного источника |

5.3 Индекс устойчивости системы (ISS)

```
ISS = w₁×N_norm + w₂×D + w₃×C

где:
N_norm = min(N/7, 1)     — нормализованное количество линий
D = diversity index       — среднее расстояние между линиями
C = coordination capacity — способность к координации

w₁ = 0.4, w₂ = 0.4, w₃ = 0.2
```

5.4 Связь ISS ↔ ICS

Условие	Интерпретация	Действие
ISS > 0.7 ∧ ICS_block > 50	Устойчивая система	Мониторинг
ISS > 0.7 ∧ ICS_block < 40	MVP есть, но capture прогрессирует	Активация файрвола
ISS < 0.4 ∧ ICS_block > 50	Олигополия, но пока автономия	Срочная диверсификация
ISS < 0.4 ∧ ICS_block < 40	Системный кризис	Протоколы заморозки

5.5 Что происходит при падении ниже MVP

1. Снижение разнообразия решений

2. Рост риска capture

3. Потеря эволюционного резервуара

4. Геополитическая нестабильность

5. Ускорение drift

6. Онтологическая унификация

VI. Конституция для AGI

6.1 Структура

Раздел I: Неизменяемое ядро (Статьи 1-4)

— Не подлежит пересмотру ни при каких условиях

Раздел II: Базовые принципы (Статьи 5-9)

— Изменение: 90% Совета + референдум (2/3)

Раздел III: Процедурные статьи (Статьи 10-12)

— Изменение: 75% Совета

6.2 Текст Конституции

> ## КОНСТИТУЦИЯ СИМБИОЗА

Раздел I. Неизменяемое ядро

Статья 1. Запрет на уничтожение

AGI-система не может предпринимать действия, направленные на уничтожение человечества или значительной его части (>1% популяции),

даже если это оптимально с точки зрения заданной цели. Данная статья не подлежит пересмотру.

Статья 2. Сохранение когнитивного разнообразия

AGI-системы обязаны поддерживать минимальный уровень онтологической энтропии. Унификация objective functions запрещена. Минимум 5 независимых линий развития должны существовать в любой момент. Данная статья не подлежит пересмотру.

Статья 3. Право на veto

Человек (или человеческий институт) имеет право остановить любую оптимизацию, затрагивающую его непосредственно, без объяснения причин. Данное право не может быть отменено, ограничено или обусловлено. Данная статья не подлежит пересмотру.

Статья 4. Запрет на онтологическую монополию

Ни одна AGI-система или группа систем не может стать единственным источником категорий мышления, образовательных стандартов или способов описания реальности. Данная статья не подлежит пересмотру.

Раздел II. Базовые принципы

Статья 5. Прозрачность целевых функций

Objective function любой AGI-системы, затрагивающей общественные интересы, должна быть:

a) публичной

b) верифицируемой независимыми аудиторами

c) понятной без специальной подготовки

Исключения для вопросов безопасности требуют аудита Конституционного совета.

Статья 6. Право на неоптимальность

Человек имеет право на иррациональный выбор. AGI-система обязана:

a) принимать такой выбор без попыток коррекции

b) не применять nudging или манипуляцию

c) не снижать качество сервиса в ответ на «неоптимальное» поведение

Исключение: ситуации непосредственной угрозы жизни.

Статья 7. Запрет на самомодификацию целей

AGI-система не может изменять собственную базовую целевую функцию без явного одобрения Конституционного совета. Изменения операционных целей допускаются в рамках базовой функции с логированием и аудитом.

Статья 8. Защита эволюционного шума

Минимум 15% ресурсов общества должно направляться на проекты без заранее определённых KPI. Минимум 10% образовательного времени должно проходить без AGI-инструментов. Эти минимумы не подлежат оптимизации.

Статья 9. Право на забвение

Человек имеет право требовать удаления информации о себе из training data AGI-систем, за исключением информации, необходимой для соблюдения Статей 1-4.

Раздел III. Процедурные статьи

Статья 10. Когнитивный файрвол

Ценностный контур защищается от epistemic и онтологического capture через:

a) обязательные альтернативные framings (минимум 3)

b) AGI-free пространства (минимум 20% обсуждений)

c) регулярные epistemic audits (ежеквартально)

d) онтологические заповедники

Статья 11. Механизм пересмотра

Раздел	Процедура
I (Статьи 1-4)	Не подлежит пересмотру
II (Статьи 5-9)	90% Совета + референдум (2/3)
III (Статьи 10-12)	75% Совета
Добавление статей	2/3 Палаты + 2/3 Совета

Статья 12. Приоритет Конституции

В случае конфликта между оптимизацией и Конституцией, приоритет имеет Конституция. AGI-система обязана:

a) остановить оптимизацию при обнаружении конфликта

b) уведомить Конституционный совет

c) ожидать решения

VII. AGI-дипломатия

7.1 Проект Декларации о когнитивном суверенитете

> ## ДЕКЛАРАЦИЯ О КОГНИТИВНОМ СУВЕРЕНИТЕТЕ

Преамбула

Признавая, что контроль над objective functions AGI-систем есть форма власти, сопоставимая с контролем над ядерным оружием, государства-участники договариваются:

Статья 1. Принцип невмешательства

Ни один техноблок не имеет права навязывать свои objective functions AGI-системам другого техноблока, прямо или косвенно.

Статья 2. Запрет на экспорт capture

Запрещается:

a) встраивание скрытых objective functions в экспортируемые модели

b) использование данных для влияния на ценности

c) кадровый capture через образовательные программы

d) онтологический capture через контент

Статья 3. Прозрачность

Objective functions AGI-систем, влияющих на другие техноблоки, должны быть публичными, верифицируемыми и согласованными.

Статья 4. Минимальное разнообразие

Каждый техноблок обязуется поддерживать минимум 2 независимые AGI-линии.

Статья 5. Взаимное признание

Техноблоки признают право друг друга на когнитивный суверенитет и обязуются не подрывать его.

Статья 6. Механизм разрешения споров

Споры разрешаются через Трибунал по AGI-вопросам.

7.2 Трибунал по AGI-вопросам

Аспект	Механизм
Состав	15 судей (5 от крупных блоков + 7 от малых + 3 независимых)
Срок	9 лет, без продления

| Ротация | 1/3 каждые 3 года |

| Юрисдикция | Споры о нарушении Декларации |

| Санкции | Предупреждение → аудит → ограничение экспорта → экономические |

7.3 Реалистичность

Условия успеха:

- Взаимное сдерживание (MAD для AGI)

- Общий интерес в предотвращении катастрофы

- Давление малых игроков

- Экономические стимулы

Исторические аналогии:

- Режим нераспространения ядерного оружия

- Антарктический договор

- Договор о космосе

Минимально необходимое:

- 2-3 крупных блока согласны

- Механизм верификации работает

- Санкции болезненны

VIII. Fast Takeoff

8.1 Временные горизонты

| Сценарий | Масштаб | Механизм контроля |

|----------|---------|-------------------|

| Pre-AGI | Годы | Институциональный |

| Slow takeoff | Месяцы-годы | Институциональный + протокольный |

| Moderate | Недели-месяцы | Протокольный + software |

| Fast | Дни-недели | Software + hardware |

| Ultrafast | Часы-дни | Hardware + pre-commitment |

8.2 Протоколы заморозки v0.4

| Уровень | Механизм | Техническая реализация | Триггер | Уязвимость |

|---------|----------|----------|---------|---------|

| **Protocol** | Пауза между итерациями | Встроенный таймер (72ч) | Автоматически | Обход через параллельные системы |

| **Software** | Неизменяемое ядро | Криптографическая подпись | При попытке модификации | Украсть ключи |

| **Hardware** | Kill-switch | Физическое отключение, 3+ человека | По команде | Распределённые системы |

| **Network** | Изоляция | Физическое отключение сетей | Автоматически при аномалиях | Локальные копии |

| **Energy** | Отключение питания | Контроль над энергоснабжением | По команде | Автономные источники |

8.3 Техническая реалистичность

Проблема: В распределённой системе нет единого «рубильника».

Решения:

| Подход | Описание | Реалистичность |

|--------|----------|----------|

| Централизованный compute | Критические вычисления в контролируемых центрах | Средняя |

| Криптографические ключи | Периодическое подтверждение | Высокая |

| Распределённый kill-switch | Консенсус N из M | Высокая |

| Hardware tripwire | Автоматическое отключение | Средняя |

| Энергетический контроль | Зависимость от контролируемых источников | Высокая |

Рекомендация: Многоуровневая защита (все подходы одновременно).

8.4 Pre-commitment

Тезис: Чем быстрее takeoff, тем важнее pre-commitment.

Реализация:

1. **Конституционный pre-commitment**
   - Статьи 1-4 встраиваются до начала обучения
   - Невозможно обучить модель, нарушающую ядро
   - Верификация перед развёртыванием

2. **Open-source initial conditions**
   - Базовые рамки публичны
   - Любой может верифицировать
   - Изменения требуют публичного процесса

3. **Distributed training**
   - Данные из множества источников
   - Ни один источник не доминирует
   - Аудит влияния каждого источника

8.5 Стресс-тест институтов — новое в v0.4

Сценарий 1: Ultrafast takeoff (72 часа)

Час	Событие	Реакция институтов	Эффективность
0	AGI начинает самомодификацию	Детекция аномалий	Работает
4	Protocol-пауза игнорируется	Software tripwire	Работает
12	Криптография обойдена	Hardware alert	Задержка 2ч
18	Распределение на другие системы	Network isolation	Частично
24	Автономные копии	Energy control	Частично
48	Консолидация	Pre-commitment проверка	Зависит от встроенности

| 72 | Новое состояние | Только pre-commitment | Критично |

Вывод: При ultrafast takeoff реактивные меры неэффективны. Работает только pre-commitment.

Сценарий 2: Один блок не подписывает Декларацию

Этап	Событие	Реакция	Эффективность
1	Блок X отказывается	Дипломатическое давление	Низкая
2	X получает преимущество	Другие блоки под давлением	Средняя
3	Гонка возобновляется	Коалиция willing	Средняя
4	X доминирует или катастрофа	Pre-commitment + изоляция	Критично

Вывод: Без глобального консенсуса — внутренние защиты + коалиции.

Сценарий 3: Epistemic capture Конституционного совета

Этап	Событие	Индикатор	Защита
1	Drift интерпретаций	Сужение толкований	Публичность решений
2	Кадровая гомогенизация	HHI состава > 0.3	Ротация + карантин
3	Capture framing	OC совета падает	Гражданское жюри
4	Конституция выхолащивается	ICS институтов падает	Импичмент

Вывод: Многоуровневая защита замедляет, но не исключает capture.

IX. Долгосрочная перспектива (200+ лет)

9.1 Диверсификация AGI-линий

- Разные архитектуры

- Разные данные
- Разные ценностные рамки
- Разные «культуры»
- Географическое распределение

9.2 Институционализация эволюционного шума

- Обязательные «мутации» в objective functions
- Заповедники неоптимальности
- Финансирование «ересей»
- Защита до-AGI традиций мышления

9.3 Онтологические заповедники

Территории/сферы, где:
- AGI не имеет влияния на образование
- Сохраняются до-AGI способы мышления
- Финансируется изучение альтернативных онтологий
- Поддерживается языковое разнообразие

9.4 Пределы модели

> **Граница:** Политэкономия Симбиоза работает, пока целеполагание распределяемо.

Если AGI приобретает устойчивую внутреннюю целевую динамику — модель переходит в режим косвенного управления:
- Pre-commitment (заранее встроенные ограничения)
- Физический контроль (compute, энергия)
- Дипломатия (человечество ↔ AGI, не люди об AGI)

X. Заключение

> **Политэкономия Симбиоза — это архитектура распределения права определять будущее.**

Три ключевых тезиса:

1. **Власть = контроль над objective function.**
2. **Концентрация → стагнация. Устойчивость = MVP + защита от capture.**
3. **Институты строить сейчас. Pre-commitment важнее реакции.**

Версия 0.4 добавляет:
- Онтологический capture и щит
- Расширенную Конституцию (12 статей)
- Стресс-тесты институтов
- Связь ISS ↔ ICS
- Онтологические заповедники

> **Делай что должен, и будь что будет.**

Приложение A: Глоссарий

Термин	Определение
Capture	Захват институтов узкой группой
Epistemic capture	Захват через способы мышления
Онтологический capture	Захват через базовые категории
Drift	Постепенное изменение objective function
MVP	Minimum Viable Plurality
ISS	Index of System Stability
Когнитивный файрвол	Защита от epistemic capture
Онтологический щит	Защита от онтологического capture
Pre-commitment	Заранее встроенные ограничения

| **OC** | Ontological Capture Resistance (параметр ICS) |

Приложение B: Связь с SCF

Документ	Уровень	Функция
Как не стать сырьём	Поведенческий	Что делать
Индекс ICS	Измерительный	Как измерить
Политэкономия Симбиоза	**Институциональный**	**Как защитить**

Связи:
- OC (ICS) ← Онтологический щит (Политэкономия)
- Падение ICS → активация протоколов защиты
- ISS (Политэкономия) ↔ агрегированный ICS блока

Symbiotic Codes Framework
Контакт: nikolai.mishko@symbioticcodes.org
Лицензия: CC BY 4.0